than the single descriptor approach. Subsequently, researchers began to use profiles for multiple windows. There could be two, three, four windows where the members of the family could agree on content. Sometimes, a profile was not built explicitly but rather was maintained as a collection of the instances across the known or alleged family members of the conserved region under consideration.--

Please replace the paragraph as it appears on page 13, lines 5-12 with the following rewritten paragraph:

--In step 210, the sequence threshold, K, is set. It is possible to set K=|T|, which is the number of sequences in the training set. In actuality, it has proven beneficial to assign a small starting value to K that is a fraction of the number of sequences in T. Experiments have shown that a starting value of K=|T|/b with b=4 or 5 is a good choice across many data sets. Note that the smaller the value of b, the higher the redundancy of the composite descriptor will be. The selection of K also can depend on how conserved, or similar, the family members are. If the family members are well conserved, then K can be higher; if the family members are not well conserved, then K can be lower.--

Please replace the paragraph as it appears on page 26, line 23, through page 27, line 5 with the following rewritten paragraph:

--Note that the 5 hits SYV_FUGRU, GTT1_RAT, GTT1_MOUSE, SYEP_HUMAN and GTH4_MAIZE are clearly separated from the 11 top scoring sequences. They however obtained scores which were above threshold and thus are studied in more detail. In all 5 cases, one or more sizeable regions that were shared with one or more members of the PS50040 collection were discovered. The Clustal-W alignment of EF1G_XENLA and the N-terminus of SYV_FUGRU, a valyl-trna synthetase from Fugu rubripes, are shown in Table 1 below. Table 1 shows a Clustal-W alignment of EF1G_XENLA and the N-terminus of SYV_FUGRU, and this shows a strong similarity. As can be seen, the similarity among these two sequences is pretty extended and the Clustal-W score for the shown alignment equaled 462.--

Please replace Table 1 as it appears on page 28 with the following rewritten Table 1:

--Table 1

```
EF1G_XENLA       (SEQ ID NO 1) MAGGTLYTYPDNWRAYKPLIAAQYSGFPIKVASSAPEFQFGVTNKTPEFLKKFPLGKVPAL
SYV_FUGRU_piece  (SEQ ID NO 2) MA--TLYVSP-----------HLDDFRSLLALVAAEY----------
                                ** ***. *            : ..*   :* *.*:

EF1G_XENLA       (SEQ ID NO 3) FEGKDGFCLFESSAIAHYVGNDELRGTTRLHQAQVIQWVSFSDSHIVPPASAWVFPTLGI
SYV_FUGRU_piece  (SEQ ID NO 4) -------C----------GNAKQ-------QSQVWQWLSFADNELTPVSCAVVFPLMGM
                                *          ** :        *:** **:**:*..:.* :.* *** :*:

EF1G_XENLA       (SEQ ID NO 5) MQYNKQATEQAKEGIKTVLGVLDSHLQTRTFLVGERITLADITVTCSLLWLYKQVLEPSF
SYV_FUGRU_piece  (SEQ ID NO 6) TGLDKKIQQNSRVELMRVLKVLDQALEPRTFLVGESITLADMAVAMAVLLPFKYVLEPSD
                                  :*:  :::: : ** ***. *:.******* *****::*: ::*  :* *****

EF1G_XENLA       (SEQ ID NO 7) RQPFGNVTRWFVTCVNQPEFRAVLGEVKLCDKMAQFDAKKFAEMQPKKETPKKEKPAKEP
SYV_FUGRU_piece  (SEQ ID NO 8) RNVLMNVTRWFTTCINQPEFLKVLGKISLCEKMVPVTAKTSTEEAAAVH-PDAAALNGPP
                               *: : ******.**:*****  ***::.**:**. . **. :*  .  . *.       *

EF1G_XENLA       (SEQ ID NO 9) KKEKEEKKKAAPTPAPAPEDDLDESEKALAAEPKSKDPYAHLP-KSSFIMDEFKRKYSNE
SYV_FUGRU_piece  (SEQ ID NO 10)KTEAQLKKEAKKREKLEKFQQKKEMEAKKKMQPVAEKKAKPEKRELGVITYDIPTPSGEK
                               *.* : **:*          :: .* *     :* ::.       : ..*  ::    .::

EF1G_XENLA       (SEQ ID NO 11)DTLTVALPYFW-EHFDKEGWSIWYAEY-KFPEELTQAFMSCNLITGMFQR-LDKLRKTGF
SYV_FUGRU_piece  (SEQ ID NO 12)KDVVSPLPDSYSPQYVEAAWYPWWEKQGFFKPEFGRKSIGEQNPRGIFMMCIPPPNVTGS
                               . :. .**  :  :: : .*  *: :   *  *: :  :. :   *:*   :   . **

EF1G_XENLA       (SEQ ID NO 13)ASVILFGTNNNSSISGVWV-FRGQDLAFTLSED-----WQIDYESYNWRKLDSGSEEC--
SYV_FUGRU_piece  (SEQ ID NO 14)LHLGHALTNAIQDTLTRWHRMRGETTLWNPGCDHAGIATQVVVEKKLMREKGTSRHDLGR
                                :    **  ..    *  :**:   :. . *      *:  *.   *: .:. .:

EF1G_XENLA       (SEQ ID NO 15)KTLVKEYFAWEGE--------FKNVGKPFNQG-KIFK----------
SYV_FUGRU_piece  (SEQ ID NO 16)EKFIEEVWKWKNEKGDRIYHQLKKLGSSLDWDRACFTMDPKLSYAVQEAFIRMHDEGVIY
                               :.:::* : *:.*        :*::*..:: .   *.
```

C4

Please replace Table 2 as it appears on page 29 with the following rewritten Table 2:

--Table 2

```
GTT1_MOUSE (SEQ ID NO 17) -VLELYLDLLSQPCRAIYIFAKKNNIPFQMHTVELRKGEHLSDAFARVNPMKKVPAMM-D
GTT1_RAT   (SEQ ID NO 18) -VLELYLDLLSQPCRAIYIFAKKNNIPFQMHTVELRKGEHLSDAFAQVNPMKKVPAMK-D
EF1G_ARTSA (SEQ ID NO 19) VAGKLYTYPENFRAFKALIAAQYSGAKLEIAKSFVFGETNKSDAFLKSFPLGKVPAFESA
                           . :**    .  .    * *: ..  ::: .  :    : **** :  *: ****:

GTT1_MOUSE (SEQ ID NO 20) GGFTLCESVAILLYLAHK-----------------------YKVPDHWYPQDLQARARV
GTT1_RAT   (SEQ ID NO 21) GGFTLCESVAILLYLAHK-----------------------YKVPDHWYPQDLQARARV
EF1G_ARTSA (SEQ ID NO 22) DGHCIAESNAIAYYVANETLRGSSDLEKAQIIQWMTFADTEILPASCTWVFPVLGIMQFN
                           .*. :.** **  *:*::                              ..  *    *

GTT1_MOUSE (SEQ ID NO 23) DEYLAWQHTGLRRSCLRALWHKVMFPVFLGEQIPPETLAATLAELDVNLQVLEDKFLQDK
GTT1_RAT   (SEQ ID NO 24) DEYLAWQHTTLRRSCLRTLWHKVMFPVFLGEQIRPEMLAATLADLDVNVQVLEDQFLQDK
EF1G_ARTSA (SEQ ID NO 25) KQATARAKEDIDKALQALDDHLLTRTYLVGERITLADIVVTCTLLHLYQHVLDEAFRKSY
                           .:  *  :  : ::      * :  . ::**:*    :..* : *.:  :**:: * :.

GTT1_MOUSE (SEQ ID NO 26) DFLVGPHISLADLVAITELMHPVGGGCPVFEGHPRLAAWYQRVEAAVGKDLFREAHEVIL
GTT1_RAT   (SEQ ID NO 27) DFLVGPHISLADVVAITELMHPVGGGCPVFEGRPRLAAWYRRVEAAVGKDLFLEAHEVIL
EF1G_ARTSA (SEQ ID NO 28) VNTNRWFITLINQKQVKAVIGDFKLCEKAGEFDP---KKYAEFQAAIGSGEKKKTEKAPK
                                 .*:* :   :. ::  .     . *  *     * ..:**:*..   ::.:.

GTT1_MOUSE (SEQ ID NO 29) KVKDCPPADLIIKQKLMPRVLTMIQ-----------------------------------
GTT1_RAT   (SEQ ID NO 30) KVRDCPPADPVIKQKLMPRVLTMIQ-----------------------------------
EF1G_ARTSA (SEQ ID NO 31) AVKAKPEKKEVPKKEQEEPADAAEEALAAEPKSKDPFDEMPKGTFNMDDFKRFYSNNEET
                           *:  *  . : *::    . :  :

GTT1_MOUSE (SEQ ID NO 32) ------------------------------------------------------------
GTT1_RAT   (SEQ ID NO 33) ------------------------------------------------------------
EF1G_ARTSA (SEQ ID NO 34) KSIPYFWEKFDKENYSIWYSEYKYQDELAKVYMSCNLITGMFQRIEKMRKQAFASVCVFG


GTT1_MOUSE (SEQ ID NO 35) ------------------------------------------------------------
GTT1_RAT   (SEQ ID NO 36) ------------------------------------------------------------
EF1G_ARTSA (SEQ ID NO 37) EDNDSSISGIWVWRGQDLAFKLSPDWQIDYESYDWKKLDPDAQETKDLVTQYFTWTGTDK


GTT1_MOUSE (SEQ ID NO 38) ------------
GTT1_RAT   (SEQ ID NO 39) ------------
EF1G_ARTSA (SEQ ID NO 40) QGRKFNQGKIFK
```

--

Please replace Table 3 as it appears on page 29 with the following rewritten Table 3:

--Table 3

```
EF1G_CAEEL (100-243) (SEQ ID NO 41)----NFD---KKTVEQYK--NELNGQLQVLDRVLVKKTYLVGERLSLADVSVALDLLPAF
SYEP_HUMAN (1-180)   (SEQ ID NO 42)MEHTEIDHWLEFSATKLSSCDSFTSTINELNHCLSLRTYLVGNSLSLADLCVWATLKGNA
                                     ::*   : :. : .  :.:.. :: *:: *  :*****: *****:.*   *

EF1G_CAEEL (100-243) (SEQ ID NO 43)QYVLDANARKSIVNVTRWFRTVVNQPAVKEV--LGEVSLASS-VA-QFNQ--AKFTELS-
SYEP_HUMAN (1-180)   (SEQ ID NO 44)AWQEQLKQKKAPVHVKRWFGFLEAQQAFQSVGTKWDVSTTKARVAPEKKQDVGKFVELPG
                                       :  : : :*: *:*.***  :  * *.:.*    :** :.: ** : :*  .**.**.

EF1G_CAEEL (100-243) (SEQ ID NO 45)---AKVAKSAPKAEKPKKEAKPAAAA--AQP-----E-------DD-EPKEEKS-KDP--
SYEP_HUMAN (1-180)   (SEQ ID NO 46)AEMGKVTVRFPPEASGYLHIGHAKAALLNQHYQVNFKGKLIMRFDDTNPEKEKEDFEKVI
                                     .**:   *   .   .   * **   *      :       ** :*::**.  :
```

Please replace Table 4 as it appears on page 30 with the following rewritten Table 4:

--Table 4

```
EF1G_RABIT (SEQ ID NO 47) MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA
GTH4_MAIZE (SEQ ID NO 48) -ATPAVKVYGWAISPFVSRALLALEEAGVDYELVPMSRQDGD-HRRPEHLARNPFGKVPV
                           *: :: .*     .* :      . * *    .*   : *: :* **.* : * ****.

EF1G_RABIT (SEQ ID NO 49) FEGDDGFCVFESNAIAYYVS----NEELRGSTPEAAAQVVQWVSFADSDIVPPAST----
GTH4_MAIZE (SEQ ID NO 50) LE-DGDLTLFESRAIARHVLRKHKPELLGGGRLEQTAMVDVWLEVEAHQLSPPAIAIVVE
                          :* *..: :***.*** :*      * * *.  * :* *  *:..   :: *** :

EF1G_RABIT (SEQ ID NO 51) WVFPTLGIMHHNKQATENAKEEVKRILGLLDAHLKTRTFLVGERVTLADITVVCTLLWLY
GTH4_MAIZE (SEQ ID NO 52) CVFAPFLGRERNQAVVDENVEKLKKVLEVYEARLATCTYLAGDFLSLADLSPF-TIMHCL
                           **..:   .:*: ..::  *::*::* : :*:* * *:*.*: ::***:: . *::

EF1G_RABIT (SEQ ID NO 53) KQVLEPSFRQAFPNTNRWFLTCINQPQFRAVLGEVKLCEKMAQFDAKKFAESQPKKDTPR
GTH4_MAIZE (SEQ ID NO 54) MATEYAALVHALPHVSAWWQGLAARP---AAN-------KVAQF--MPVGAGAPKEQE--
                            .  .:: :*:*:.. *:     :*   *.        *:***    .. . **::
```

--

Please replace the paragraph as it appears on page 31, lines 10-24 with the following rewritten paragraph:

--The collection of 804 GPCR sequences and fragments contained several classes (e.g. rhodopsin-like, secretin-like, pheromone, etc.) of proteins. In turn, each of these classes comprised several representatives. Instead of selecting representatives from each of the identified classes, the order of the sequences in this set of 804 members were randomized. Note that the contents of the sequences themselves remained unchanged, only their order of appearance was modified. For example, the 613-th sequence was now listed 4-th, the 11-th sequence now appeared in the 45-th position, and so on.

Subsequently, a training set T was formed by collecting the sequences and fragments listed in the first 80 positions, arguably a very small set if one considers the diversity of the GPCR family. Essentially, slightly less than 1/10-th of the available dataset were randomly sub-selected for the purposes of building the composite descriptor. Table 5 below contains a listing of the labels of the 80 sequences in this training set. Table 5 shows the Swiss-Prot labels of the 80 sequences in the training set for the G protein-coupled receptor experiment. The labels are listed in the order they were selected and they correspond to both sequences and sequence fragments.--

Please replace the paragraph as it appears on page 43, lines 10-14 with the following rewritten paragraph:

--The three composite descriptors were used to search the collection of 19,099 ORFs that were reported for the C. elegans genome, by the Washington University in St. Louis, School of Medicine, Genome Sequence Center, as of June 13, 1999. In all three cases, the corresponding values of $Thres_{rand}$ that were established by searching RAND-Swiss-Prot were used.--

Please replace the paragraph as it appears on page 48, lines 4-20 with the following rewritten paragraph:

--The fragments were:

```
>Y94H6A_142.g fragment (SEQ ID NO 55)
IFDNTNDLVASLLGISSITVYRKRKRIGEE
>C16C2.1 fragment (SEQ ID NO 56)
YLSGSTRAKLAESLGLSDNQVKVWFQNRRT
>F18C5.2 fragment (SEQ ID NO 57)
ISRSTAKEVATARGISEGTVYSYLAMAVEK
>Y39F10A.a fragment (SEQ ID NO 58)
LSAYTISDLAKHFNVSKIEILKIDIEGAEL
>Y48C3A.s fragment (SEQ ID NO 59)
NEVLNLNEVAKELNISKRRVYDVINVLEGL
```

and their respective top-scoring sequences from the training set of 70 helix-turn helix segments, blast scores, P and N values are:

#	C. elegans ORF	Top Scoring	Scor	P	N
1	Y94H6A_142.g	RPSF_BACSU	50	2.80E-06	1
2	C16C2.1	TER3_ECOLI	45	1.30E-05	1